Problem 4: Sports OR Politics

Shubham Haraniya
Roll No: M25CSA013
Natural Language Understanding

Abstract

In this project we did binary classification between Sports POlitics using News Category Dataset.I implemented using three ml algorithms Naive Bayes, SVM Logistic Regression with differnt feature representation Bag of words, TF IDF N Grams. Naive Bayes with simple word count achieved **96.60% accuracy** highest in all other model and representation.

1 Introduction

Methodology

The pipeline has three parts:

1. Text preprocessing and normalization
2. Feature extraction and vectorization
3. Model training and evaluation

2 Data Collection and Exploration

Dataset: News Category Dataset (HuffPost articles 2012-2022)
Source: Kaggle - News Category Dataset
Selected Categories:

- **Sports**: 5,077 articles covering various sports topics
- **Politics**: 35,602 articles (reduced to 5,077 via random sampling)

2.1 Class Distribution

The training set have **4,062 political** sample and **4,062 sports** sample (see Figure 1). The dataset is perfectly balanced as we randomly sampled from the larger Politics category to match the Sports count eliminating any class imbalance problem.

Figure 1: Train set distribution showing perfect class balance

2.2 Vocabulary Analysis

I seen most frequent term in each category after removing stopwords.

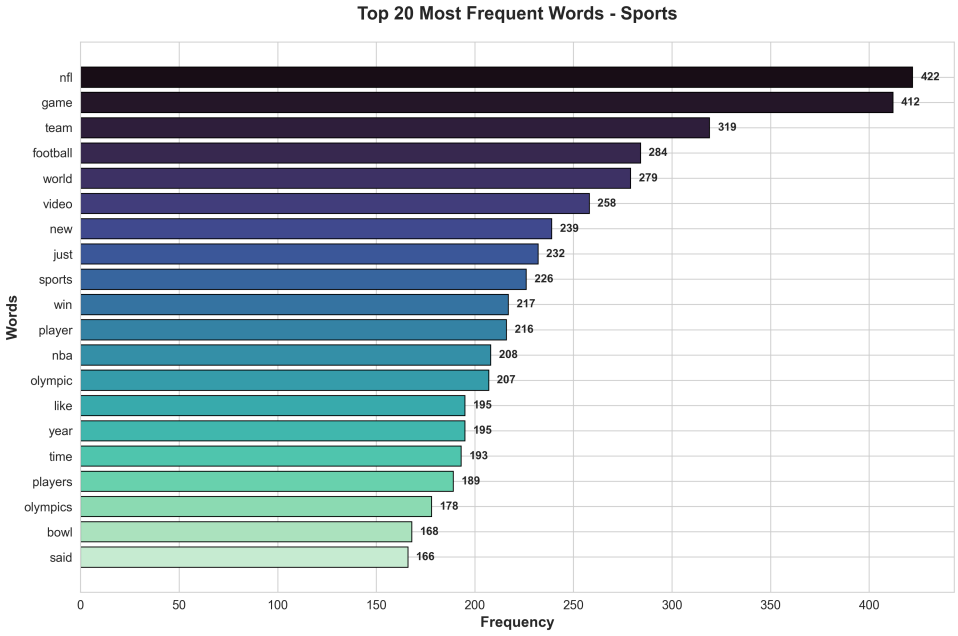

Figure 2: Top 20 words in Sports documents

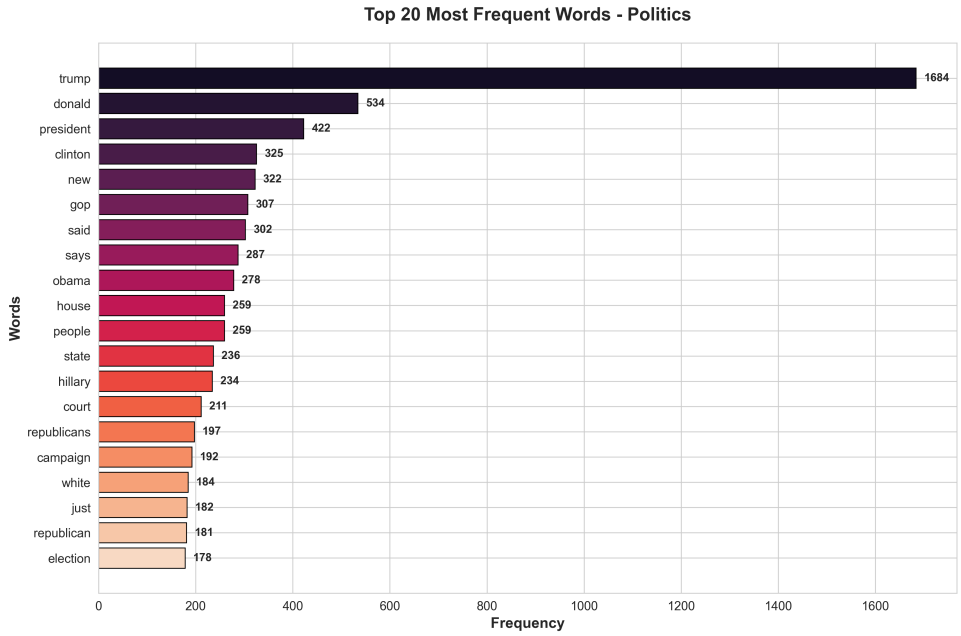

Figure 3: Top 20 words in Politics documents

Insight

Figures 2 and 3 reveal strong differnce between classes. Sports vocabulary has "nfl", "game", "team", "football" while Politics features "trump", "donald", "president", "clinton".

3 Methodology

3.1 Preprocessing

Preprocessing Pipeline

1. convert to lowercase
2. Remove stopword using NLTK English stopword list

3.2 Feature Extraction

Three Feature Extraction :

Bag of Words: Raw term frequency counts. it is straightforward but ignores word importance.
TF-IDF: Weights terms by their inverse document frequency theoretically highlighting discriminative vocabulary while put out common terms.
N grams: Captures word pair with unigram to maintain some local context. I might thought it help with phrase like gun control or playoff series

3.3 Classification Algorithms

Methodology

I tested three algorithms:
Naive Bayes: Despite independence assumption being clearly avoided in natural language this probabilistic classifier work well for text.
Support Vector Machines: Linear kernel default regularization (C=1.0). SVM genrelly achieve strong performance on high dimensional sparse data.
Logistic Regression: L2 regularization with default parameters. Provides probabilistic outputs.

4 Results

4.1 Model Comparison

Key Result

Table 1 present accuracy and F1 scores in all parts. All models got above **94% accuracy** confirming that Sports and Politics are well separated in feature space.

Feature	Algorithm	Accuracy	F1-Score
Bag of Words	**Naive Bayes**	**96.60%**	**0.97**
Bag of Words	SVM	94.44%	0.94
Bag of Words	Logistic Regression	95.62%	0.96
TF-IDF	Naive Bayes	96.21%	0.96
TF-IDF	SVM	95.86%	0.96
TF-IDF	Logistic Regression	95.22%	0.95
N-grams (1,2)	Naive Bayes	96.36%	0.96
N-grams (1,2)	SVM	96.31%	0.96
N-grams (1,2)	Logistic Regression	95.27%	0.95

Table 1: Classification performance metrics

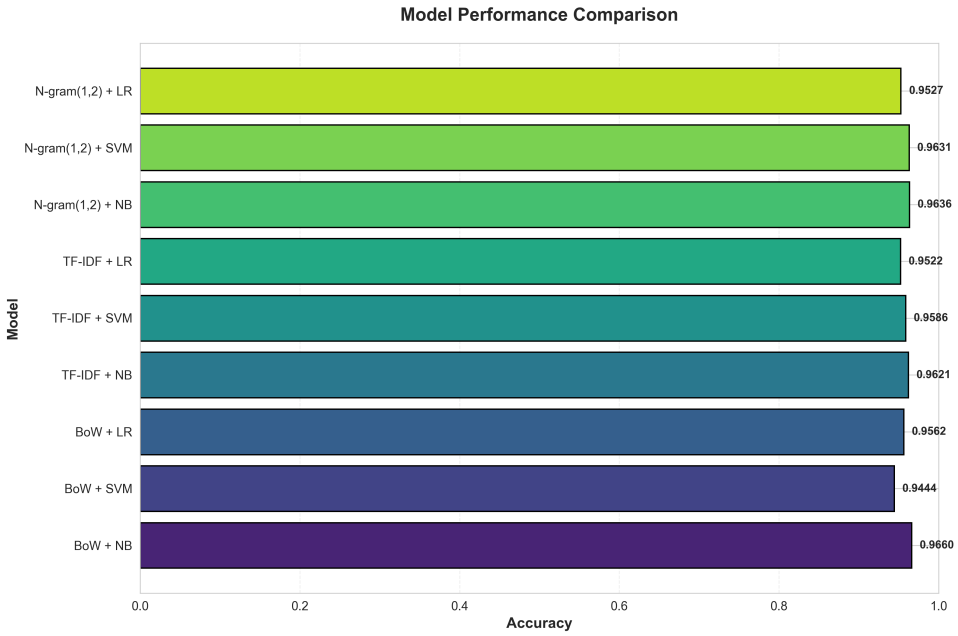

Figure 4: Accuracy comparison

4.2 Analysis

Insight

Naive Bayes with BoW: The simple configuration Naive Bayes with raw word counts got highest accuracy (96.60%). This makes sense for this task. two categories use almost differnt vocabularies, so word presence or absence provides strong signal. Naive Bayes perform well at this treating each word independent evidence and multiplying probabilities.

More complex models might try to learn complex decision boundaries that don't exist or aren not needed. When the classes are this separable complexity adds noise rather than signal.

TF-IDF Performance

TF-IDF slightly helped SVM (95.86% vs 94.44%) but actually hurt Naive Bayes (96.21% vs 96.60%). Tdown common terms probably does not matter when the vocabularie are already so diverse. For Naive Bayes specifically TF IDF breaks the generative probability model assumptions without providing benefits.

N-gram Performance

Adding bigrams showed mixed results. Naive Bayes bit improved to 96.36% while SVM improved to 96.31%. However these gains are minimal two observation:
First, dimensionality unigram vocabulary is already around 10,000 terms adding bigrams pushes this to 50,000+ features. Many bigrams appear only once or twice provide no generalizable signal.
Second task does not require context. Gun is informative not matter around words. Hockey signals Sports even it appears in hockey game or ice hockey Bigrams add complexity without capture meaningful pattern.

4.3 Error Analysis

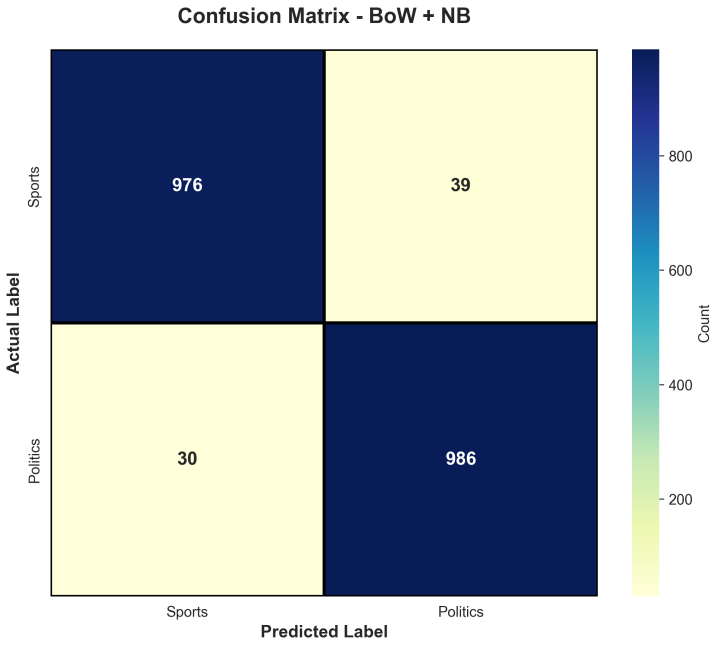

Figure 5: Confusion matrix for best model NB + BoW

Error Analysis

Out of 2,031 test documents **69 were misclassified** by the best model

- One Politic corpus used sports words (political football, touchdown for the administration). sports vocabulary fooled classifier.
- A document on sports has stadium security regulation and firearm policie at sports events combine vocabulary from both area.
- two documents were questionable even for human judgment as they contained very short quoted text.

These mistakes make sense when you consider limits of a bag of words model. since it does not understand meaning or context it struggle to tell difference between literal language and metaphorical expressions.

5 Conclusion

Key Result

For the Sports vs Politics classification task simpler approaches performed surprisingly well naive Bayes combined with Bag of Words representation achieved **96.60% accuracy**, outperforming both complex feature engineering techniques and more advanced algorithm like SVM.

This result shows key principle: model complexity should be aligned with task difficulty. when classes have highly distinctive vocabularies as is the case with Sports and Politics more elaborate methods can actually hurt performance by overfitting noise or learning not needed patterns.

The few errors made by the best performing model occurred in cases where vocabulary overlapped in unexpected ways such as metaphorical language or cross-domain topics. Addressing these errors would require a level of semantic understanding that bag of words models simply do not provide.